CPSC 440/540: Machine Learning

Empirical and Hierarchical Bayes

Winter 2023

Last Time: Beta-Bernoulli coin flipping

- Flipping a coin :
  - Prior $p(\theta)$ — $\theta \sim \text{Beta}(\alpha, \beta)$ — Uniform: $\alpha = 1, \beta = 1$
  - Likelihood $p(x \mid \theta)$ — $X \mid \theta \sim \text{Bernoulli}(\theta)$
  - Posterior $p(\theta \mid \boldsymbol{X})$ — $\theta \mid \boldsymbol{X} \sim \text{Beta}(\alpha + n_1, \beta + n_0)$ — $\text{Beta}(n_1 + 1, n_0 + 1)$
- Want to estimate $\Pr(x^3 = 1 \mid x^1 = 1, x^2 = 1)$
- MAP: becomes MLE
  - $\hat{\theta} = \text{argmax}\, p(\theta \mid x^1, x^2)$ — 1 if $\beta \leq 1$, $\frac{\alpha+1}{\alpha+\beta}$ if $\beta \geq 1$ — chooses $\hat{\theta} = 1$
  - Answer $\Pr(x^3 = 1 \mid \hat{\theta})$ — 1 if $\beta \leq 1$, $\frac{\alpha+1}{\alpha+\beta}$ if $\beta \geq 1$ — estimate is 1
- Bayesian learning:
  - Marginalize over all possible values of θ:

$$\Pr(x^3 = 1 \mid x^1, x^2) = \int \Pr(x^3 = 1 \mid \theta)\, p(\theta \mid x^1, x^2)\, d\theta = \frac{\alpha + 2}{\alpha + \beta + 2} = \frac{3}{4}$$

Motivation: Controlling Complexity

- For many application, we need complicated models.
- But complex models can overfit.
- So what should we do?

- In CPSC 340 we saw two ways to reduce overfitting:
  - Model averaging (like in random forests).
  - Regularization (like in L2-regularized linear regression).

- Bayesian methods combine both of these.
  - Average over "models", weighted by posterior (which includes regularizer).
    - Recall that the regularizer corresponds to the negative logarithm of the prior.
  - This can allow you fit extremely complicated models without overfitting.

MAP vs Bayes for Categorical-Dirichlet

- MAP (regularized optimization) approach maximizes over parameters:

$$\hat{\theta} \in \arg\max_{\theta} \{ p(\theta \mid X) \}$$
$$= \arg\max_{\theta} \{ p(X \mid \theta) p(\theta) \} \quad \text{(Bayes' rule)}$$
$$p(x = c \mid \hat{\theta}) = \hat{\theta}_c$$

(I'm not explicitly including the conditioning on the hyper-parameters α)

- Bayesian approach predicts by integrating over possible parameters:

$$p(x = c \mid X) = \int_{\theta_1} \int_{\theta_2} \cdots \int_{\theta_K} p(x = c, \theta \mid X) \, d\theta_K \, d\theta_{K-1} \cdots d\theta_1 \quad \text{(marg. rule)}$$
$$= \int_{\theta_1} \int_{\theta_2} \cdots \int_{\theta_K} p(x = c \mid \theta, X) \, p(\theta \mid X) \, d\theta_K \, d\theta_{K-1} \cdots d\theta_1 \quad \text{(product rule)}$$
$$= \int_{\theta_1} \int_{\theta_2} \cdots \int_{\theta_K} \theta_c \, p(\theta \mid X) \, d\theta_K \, d\theta_{K-1} \cdots d\theta_1$$

$p(\theta \mid X) \propto p(X \mid \theta) p(\theta)$

(independence of data given parameters)

$\mathbb{E}[\theta_c]$ (mean of Dirichlet posterior)

- Considers all possible θ, and weights prediction by posterior for θ.
  - Posterior contains a regularizer, so this is averaging and regularizing.

$\mathbb{E}_{\theta \sim \text{posterior}}[g(\theta)] = \int g(\theta)\, p(\theta \mid X)\, d\theta$

Ingredients of Bayesian Inference (MEMORIZE)

1. Likelihood $p(X \mid \Theta)$
   - Probability of seeing data given parameters.
2. Prior $\mathrm{p}(\Theta \mid \mathrm{A})$.
   - Belief that parameters are correct before we have seen data.

things you choose

3. Posterior $\mathrm{p}(\Theta \mid \boldsymbol{X}, \mathrm{A})$.
   - Probability that parameters are correct after we have seen data.
   - MAP maximizes, but Bayesian approach uses the whole distribution.
4. Posterior predictive $p(\tilde{X} \mid \boldsymbol{X}, \mathrm{A})$ (NEW).
   - Probability of new data $\tilde{X}$ given old data $\boldsymbol{X}$, integrating over parameters.
     - Specifically, we average the likelihood of $\tilde{X}$, weighted by the posterior of θ given $\boldsymbol{X}$.
   - Bayesian approach uses this distribution for inference.

math tells you

Bayesian Approach: Discussion

- Our previous "learn then predict" approaches (MLE and MAP):
  - Optimize parameters θ (learning).
  - Do inference with the parameter estimate $\hat{\theta}$ (inference).

- Bayesian approach doesn't really have a separate "learning phase".
  - There is no optimization of the parameter θ.
  - You just skip to doing inference with the posterior predictive.
    - Consider all parameters θ.

- In practice, it often still looks like "learn then predict".
  - Characterize the form of the posterior ("learning").
  - Make predictions by doing integrals with the posterior (inference).

Bayesian Approach: Discussion

- The Bayesian approach is the optimal way to use a probabilistic model.
  - It's what the rules of probability say we should do.
  - …if you believe in your probability model (prior + likelihood).

- If the prior (or likelihood) is bad, Bayesian approach can be harmful.
  - Bayesian approach historically criticized since it requires "subjective" prior.
  - But all models are based on "subjective" assumptions, sometime hidden!

- As we see more data, Bayesian posterior concentrates on MLE.
  - MLE/MAP/Bayes usually more or less agree for large datasets.

- Real problem with the Bayesian approach is that integrals are hard.
  - Posterior and posterior predictive only have a nice form with conjugate priors.
    - Otherwise, you need to use methods like Monte Carlo or "variational" methods for inference.

Next Topic: Empirical Bayes

Learning the Prior from Data?

- How do we tune the hyper-parameters in Bayesian methods?

- Adapting our usual validation set approach:
  - Split into a training and validation set.
  - For different hyper-parameter values:
    - Compute some measure of "test error".
      - For density estimation, this could be the (negative log) posterior predictive for the validation set given the training set.
      - For supervised learning, you could make predictions on the validation set and measure validation set error.
  - Choose the hyper-parameters with the highest value.

- Advantage:
  - Directly tunes hyper-parameters to achieve good performance on new data.
- Disadvantage:
  - Optimization bias: can start to overfit to the validation set.
  - Slow! If you try 10 values for each of k hyper-parameters, there are 10^k values to try.

$X|\theta \sim \text{Bern}(\theta)$

$\theta|\alpha,\beta \sim \text{Beta}(\alpha,\beta)$

$p(X|\alpha,\beta) = \int p(X|\theta)\, p(\theta|\alpha,\beta)\, d\theta$

Learning the Prior from Data?

- Empirical Bayes:
  - Optimize the likelihood of the data given the hyper-parameters.

$$\hat{\alpha} \in \operatorname*{argmax}_{\alpha} \{p(X|\alpha)\} \equiv \operatorname*{argmax}_{\alpha} \left\{ \int p(X|\Theta)\, p(\Theta|\alpha)\, d\Theta \right\}$$

I am writing this as / integral even if there are many parameters.

marg. rule, product rule, cond. ind.

    - This is called the "marginal likelihood" or the "evidence" function.
      - It can be computed by marginalizing over parameters.
      - It's the denominator we ignore when we do MAP estimation: $p(\Theta \mid X) = \frac{p(X \mid \Theta)p(\Theta \mid \mathrm{A})}{p(X \mid \mathrm{A})}$.
    - Empirical Bayes is also called "type II maximum likelihood" or "evidence maximization".
  - This is doing MLE for the hyper-parameters.

- Advantage:
  - Fast! Might have a closed-form solution or allow using gradient descent (assuming conjugate prior).
- Disadvantage:
  - It is not directly testing the performance on new data.
  - Optimization bias: can start to overfit the marginal likelihood (could increase/decrease test performance).

Marginal Likelihood with Conjugate Priors

- Marginal likelihood has closed form when using conjugate priors.
  - It is proportional to ratio of posterior/prior normalizing constants.

- We will show this for the Bernoulli-Beta model:

$$p(X \mid \theta) = \theta^{n_1}(1-\theta)^{n_0}$$ (Likelihood)

$$p(\theta \mid \alpha, \beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{Z(\alpha,\beta)}$$ (Prior)

$$p(\theta \mid X, \alpha, \beta) = \frac{\theta^{(n_1+\alpha)-1}(1-\theta)^{(n_0+\beta)-1}}{Z(n_1+\alpha, n_0+\beta)}$$ (Posterior)

$$Z(\alpha,\beta) = \int \theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta$$ (Normalizing constant)

$$p(X \mid \alpha, \beta) = \int p(X \mid \theta)\, p(\theta \mid \alpha, \beta)\, d\theta$$ (marginal likelihood)

$$= \int \theta^{n_1}(1-\theta)^{n_0} \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{Z(\alpha,\beta)} d\theta = \frac{1}{Z(\alpha,\beta)} \underbrace{\int \theta^{(n_1+\alpha)-1}(1-\theta)^{(n_0+\beta)-1} d\theta}_{Z(n_1+\alpha, n_0+\beta)} = \frac{Z(n_1+\alpha, n_0+\beta)}{Z(\alpha,\beta)}$$

Marginal Likelihood with Conjugate Priors

- For the Bernoulli-beta model we have marginal likelihood of:

$$p(X \mid \alpha, \beta) = \frac{Z(n_1+\alpha, n_0+\beta)}{Z(\alpha,\beta)}$$

  - For other distributions the ratio might be multiplied by a constant.
    - By similar argument, posterior predictive for new data with counts $\tilde{n}_1$ and $\tilde{n}_0$ is:

$$\frac{Z(n_1+\tilde{n}_1+\alpha, n_0+\tilde{n}_0+\beta)}{Z(n_1+\alpha, n_0+\beta)}$$

- Empirical Bayes maximizes marginal likelihood in terms of α and β.
  - More useful when we have many hyper-parameters.
  - Could be used for categorical-Dirichlet model's *k* hyper-parameters.
  - In some cases, this is related to leave-one-out cross-validation.
    - The "most extreme" form of cross-validation.

Learning Principles for Predicting "0 or 1 Next?"

- Maximum likelihood:

$$\hat{\theta} \in \operatorname*{argmax}_{\theta} \{ p(X \mid \theta) \} \qquad \hat{x} \in \operatorname*{argmax}_{x} \{ p(x \mid \hat{\theta}) \}$$

- MAP: $\hat{\theta} \in \operatorname*{argmax}_{\theta} \{ p(\theta \mid X, \alpha, \beta) \}$ $\hat{x} \in \operatorname*{argmax}_{x} \{ p(x \mid \hat{\theta}) \}$

- Bayesian (no "learning"):

$$\hat{x} \in \operatorname*{argmax}_{x} \{ p(x \mid X, \alpha, \beta) \} \equiv \operatorname*{argmax}_{x} \{ \int p(\theta \mid X, \alpha, \beta) p(x \mid \theta) d\theta \}$$

- Empirical Bayes:

$$\hat{\alpha}, \hat{\beta} \in \operatorname{argmax} \{ p(X \mid \alpha, \beta) \} \qquad \hat{x} \in \operatorname{argmax} \{ p(x \mid X, \hat{\alpha}, \hat{\beta}) \}$$

Bayesian Hierarchy

- Maximum likelihood estimation can do weird things.
  - Predict zero probability for events not seen in training.
  - Pick a highly-unlikely model that exactly fits the training data.
- MAP estimation improves MLE by adding a prior on the parameters..
  - But by only using one parameter estimate this leads to sub-optimal decisions.
- Bayesian inference over parameters makes optimal decisions.
  - Avoids overfitting, and decisions follow rules of probability.
    - No optimization bias because no optimization.
  - But this relies on have a good choice of prior/hyper-parameters.
- Empirical Bayes uses data to find a good prior.
  - Tends to be less sensitive to overfitting than regular MLE.
  - But has an optimization bias: can still overfit the hyper-parameters.
  - In my experience, more likely to "just be weird" than actual overfitting.

https://www.sweetsugarbelle.com/2013/01/that-funky-bandaid-color/

Bayesian Hierarchy

- To fix empirical Bayes issues:
  - We can put a prior on the hyper-parameters.
  - Sometimes called a "hyper-prior", that has "hyper-hyper-parameters".
    - Seriously!
  - But by only using one parameter estimate this leads to sub-optimal decisions.
- So use Bayesian inference over parameters and hyper-parameters:
  - You would integrate over all values of the parameters and hyper-parameters.
    - Unfortunately, we often do not have a "conjugate hyper-prior" for the prior.
  - This avoids overfitting, but now we rely on having a good choice of hyper-prior.
- And then could consider empirical Bayes over hyper-hyper-parameters…
  - This was one the hottest ML topics before deep learning came back.

https://en.wikipedia.org/wiki/Turtles_all_the_way_down
https://www.sweetsugarbelle.com/2013/01/that-funky-bandaid-color/

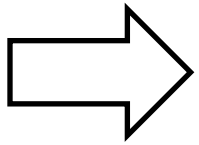
Next Topic: Hierarchical Bayes

Motivating Example: Medical Treatment

- Consider modeling probability that a medical treatment will work.
  - But this probability depends on the hospital where treatment is given.
- So we have binary examples $x^1, x^2, \ldots, x^n$.
  - We also have a number z^i saying "what population it came from".
    - This is a common non-IID setting: examples are only IID within each group.

X =

Worked?	Hospital
1	1
0	4
0	3
1	2
0	3



$p(x = 1 \mid z=2) = 0.4$

- Other examples:
  - "What are the covid proportions for different cities?"
  - "Which of my stores will sell over 100 units of product?"
  - "What proportion of users will click my ads on different websites?"

Independent Model for Each Group

- We could consider a simple independent model for each group:
  - Use a parameter θ_j for each hospital 'j'.

$$x^i | z^i \sim Ber(\theta_{z^i})$$

  - Fit each θ_j using only the data from hospital 'j'.
    - If we have 'k' hospitals, we solve 'k' IID learning problems.

- Problem: we may not have a lot of data for each hospital.
  - Can we use data from a hospital with a lot of data to learn about others?
  - Can we use data across many hospitals to learn with less data?
  - Can we say anything about a hospital with no data?

Dependencies from Using a Common Prior

- Common approach: assume the θ_j are drawn from a common prior.

$$x^i \mid z^i \sim Ber(\theta_{z^i}) \qquad \theta_j \sim Beta(\alpha, \beta)$$

- This introduces a dependency between the θ_j values.
  - For example, if $\alpha = 5$ and $\beta = 2$:
    - This is like we imagine seeing 5 extra "success" and 2 "failures" at each hospital.

- In this setting the θ_j are conditionally independent given α and β.
  - With a fixed prior, we cannot learn about one θ_j using data from another.
    - So for a new hospital, the posterior over θ_j is the prior.
- In this setting, we want to learn the hyper-parameters.

Hierarchical Bayesian Modeling

thread on Beta's conjugate prior

- Consider using a hyper-prior:

$$x^i \mid z^i \sim Ber(\theta_{z^i}) \qquad \theta_j \sim Beta(\alpha, \beta) \qquad \alpha, \beta \sim D(p, q, m)$$

(conjugate prior for beta has 3 parameters)

  - Treating hyper-parameters as random variables, can learn across groups.
- With empirical Bayes we get fixed estimates of $\tilde{\alpha}$ and $\tilde{\beta}$.
  - Learned prior gives better estimates of θ_j for groups with few examples.
  - For a new hospital, posterior would default to the learned prior.
- With hierarchical Bayes we would integrate over the θ_js, α, and β.
  - "Very Bayesian" to handle the unknown parameters/hyper-parameters.
  - Hierarchical models almost always need approximations like Monte Carlo.

Discussion of Hierarchical Bayes

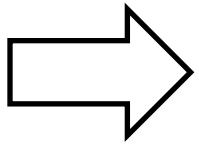
- Many practitioners really like Bayesian models.
  - "Gosh darn, I love Bayesian ensemble methods!"
    - From a domain expert I was collaborating with.
  - Domain expertise can be incorporated into the design of [hyper-]priors.
  - Can model various ways your data may not be IID.
  - We will see some more Bayes tricks.
- Advantage is the nice mathematically framework:
  - Write out all your prior knowledge of relationships between variables.
  - Integrate over variables you do not know.
- Disadvantages:
  - It can be hard to exactly encode your prior beliefs.
  - The integrals get ugly very quickly (there is no "automatic integration").

Motivating Example: Medical Treatment

- Consider modeling probability that a medical treatment will work.
  - But this probability depends on the hospital where treatment is given.
- So we have binary examples $x^1, x^2, \ldots, x^n$.
  - We also have a number z^i saying "what population it came from".
    - This is a common non-IID setting: examples are only IID within each group.

X =

Worked?	Hospital
1	1
0	4
0	3
1	2
0	3



$p(x = 1 \mid z=2) = 0.4$

- Other examples:
  - "What are the covid proportions for different cities?"
  - "Which of my stores will sell over 100 units of product?"
  - "What proportion of users will click my ads on different websites?"

Independent Model for Each Group

- We could consider a simple independent model for each group:
  - Use a parameter θ_j for each hospital 'j'.

$$x^i | z^i \sim Ber(\theta_{z^i})$$

  - Fit each θ_j using only the data from hospital *j*.
    - If we have *k* hospitals, we solve *k* IID learning problems.

- Problem: we may not have a lot of data for each hospital.
  - Can we use data from a hospital with a lot of data to learn about others?
  - Can we use data across many hospitals to learn with less data?
  - Can we say *anything* about a hospital with no data?

Dependencies from Using a Common Prior

- Common approach: assume the θ_j are drawn from a common prior.

$$x^i \mid z^i \sim Ber(\theta_{z^i}) \qquad \theta_j \sim Beta(\alpha, \beta)$$

- This introduces a dependency between the θ_j values.
  - For example, if $\alpha = 5$ and $\beta = 2$:
    - This is like we imagine seeing 5 extra "success" and 2 "failures" at each hospital.

- In this setting the θ_j are conditionally independent given α and β.
  - With a fixed prior, we cannot learn about one θ_j using data from another.
    - So for a new hospital, the posterior over θ_j is the prior.
- In this setting, we want to learn the hyper-parameters.

Hierarchical Bayesian Modeling

thread on Beta's conjugate prior

- Consider using a hyper-prior:

$$x^i \mid z^i \sim Ber(\theta_{z^i}) \qquad \theta_j \sim Beta(\alpha, \beta) \qquad \alpha, \beta \sim D(p, q, m)$$

(conjugate prior for beta has 3 parameters)

  - Treating hyper-parameters as random variables, can learn across groups.
- With empirical Bayes we get fixed estimates of $\tilde{\alpha}$ and $\tilde{\beta}$.
  - Learned prior gives better estimates of θ_j for groups with few examples.
  - For a new hospital, posterior would default to the learned prior.
- With hierarchical Bayes we would integrate over the θ_js, α, and β.
  - "Very Bayesian" to handle the unknown parameters/hyper-parameters.
  - Hierarchical models almost always need approximations like Monte Carlo.

Discussion of Hierarchical Bayes

- Many practitioners really like Bayesian models.
  - "Gosh darn, I love Bayesian ensemble methods!"
    - From a domain expert Mark was collaborating with.
  - Domain expertise can be incorporated into the design of [hyper-]priors.
  - Can model various ways your data may not be IID.
  - We will see some more Bayes tricks.
- Advantage is the nice mathematical framework:
  - Write out all your prior knowledge of relationships between variables.
  - Integrate over variables you do not know.
- Disadvantages:
  - It can be hard to exactly encode your prior beliefs.
  - The integrals get ugly very quickly (there is no "automatic integration").

bonus!

METHODS article

Front. Environ. Sci., 09 March 2021 | https://doi.org/10.3389/fenvs.2021.491636

Evaluating the Benefits of Bayesian Hierarchical Methods for Analyzing Heterogeneous Environmental Datasets: A Case Study of Marine Organic Carbon Fluxes

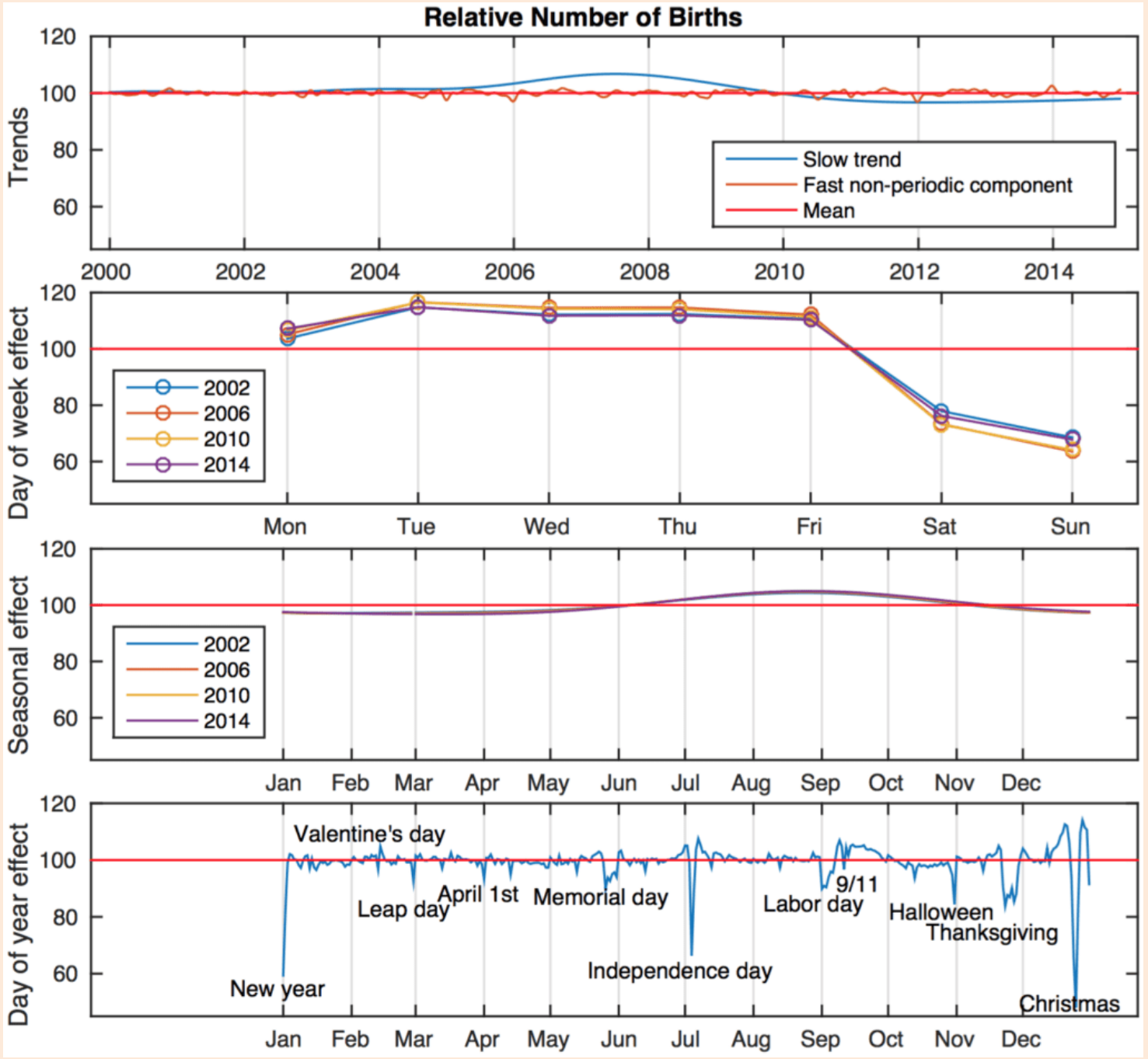
observations from 407 sampling locations spanning eight biomes across the global ocean. We fit a global scale Bayesian hierarchical model that describes the vertical profile of organic carbon flux with depth. Profile parameters within a particular biome are assumed to share a common deviation from the global mean profile. Individual station-level parameters are then modeled as deviations from the common biome-level profile. The hierarchical approach is shown to have several benefits over simpler and more common data aggregation methods. First, the hierarchical approach avoids statistical complexities introduced due to unbalanced sampling and allows for flexible incorporation of spatial heterogeneitites in model parameters. Second, the hierarchical approach uses the whole dataset simultaneously to fit the model parameters which shares information across datasets and reduces the uncertainty up to 95% in individual profiles. Third, the Bayesian approach incorporates prior scientific information about model parameters; for example, the non-negativity of chemical concentrations or mass-balance, which we apply here. We explicitly quantify each of these properties in turn. We emphasize the generality of the hierarchical Bayesian approach for diverse environmental applications and its increasing feasibility for large datasets due to recent developments in Markov Chain Monte Carlo algorithms and easy-to-use high-level software implementations.

We will cover MCMC later

Parameter(s) for each group
→ Groups with many samples inform other groups
→ All data is used.
→ Prior and hyper-prior can have domain knowledge
→ Monte Carlo used to approximate ugly integrals

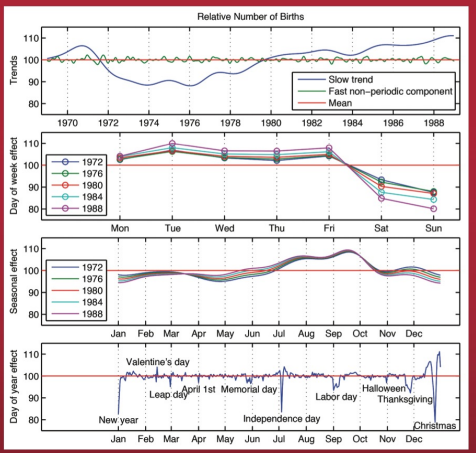
https://www.frontiersin.org/articles/10.3389/fenvs.2021.491636/full

bonus!



Bayesian Data Analysis

Third Edition



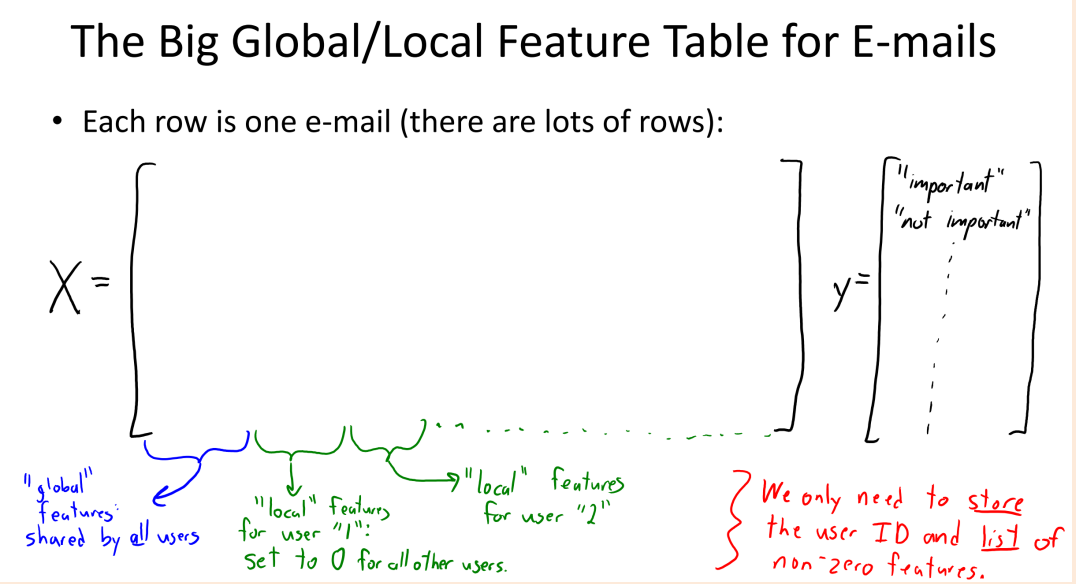
Andrew Gelman, John B. Carlin, Hal S. Stern,
David B. Dunson, Aki Vehtari, and Donald B. Rubin

stat.columbia.edu/~gelman/book/

bonus!

Discussion of Hierarchical Bayes

- "We finally have an elegant mathematical way to do…"
  - Frequently used as a justification for hierarchical Bayesian methods.
  - We will see some influential and/or neat examples later in the course.
- But often you can find a simple less-elegant solution:
  - 340 slide giving features addressing similar issues to hospital example.
    - Just features and gradients, no hyper-priors or integrals.



Summary

- Marginal likelihood:
  - Probability of data given hyper-parameters (integrating over parameters).
- Empirical Bayes ("type II MLE" or "evidence maximization").
  - Tune hyper-parameters by optimizing marginal likelihood.
  - Can be used to cheaply tune a huge number of hyper-parameters.
    - If you can efficiently do/approximate the integrals.
- Hyper-priors:
  - Putting a prior on the prior.
  - Often needed to make empirical Bayes work, or in hierarchical Bayes.
- Hierarchical Bayes:
  - Building models with multiple levels of priors.
  - Often allows learning in non-standard scenarios.
    - We considered the case of non-IID grouped data.

- Next Time: everyone's favourite loss to take the gradient of.